GESP 3B / SONAR: An Introspection-First Protocol for Behavioral Grounding and Structured Emergence in Large Language Models

Micah Hughes — Independent Researcher — 2026

Abstract

We present GESP 3B/SONAR, a lightweight inference-time protocol that restructures generation in large language models to produce behaviorally grounded outputs without modifying model weights or requiring internal access. The core contribution is a pass-wise sensing loop — detect structure before naming, cite boundary evidence before committing, surface uncertainty before releasing output — that shows evidence consistent with reduction in premature confident commitment across text and vision-language tasks. We introduce a secondary contribution: the SONAR reframing, which proposes that refusal and guardrail dynamics produce measurable resistance gradients that can be treated as proto-sensory signals for boundary mapping rather than simple stopping conditions. We report a 0.976 composite score under the protocol's internal rubric across six behavioral metrics in a single-stage memoryless run under strict constraints, provide a full replication package including canonical seeds, diagnostic gates, ablation protocol, and adversarial reviewer kit, and demonstrate measurable self-characterization behavior under constrained metalinguistic prompting. The protocol is architecture-agnostic, adds minimal inference overhead, and is designed for independent falsification. We clearly demarcate demonstrated results from proposed extensions, with implications for grounding, uncertainty quantification, hallucination reduction, and inference-time alignment research.

Keywords: *inference-time grounding, boundary gradient sensing, behavioral uncertainty surfacing, evidence-anchored generation, self-characterization, hallucination reduction, refusal dynamics, constrained generation, introspection-first protocol, pass-wise generation*

---

1. Introduction

The central reliability problem in deployed large language models is not capability — it is premature commitment. Models generate answers at uniform apparent confidence regardless of whether the underlying evidence supports the claim. The mechanism that produces a correct and well-grounded statement is structurally identical to the mechanism that produces a confident confabulation. From the output alone, and often from the generation process itself, the two are indistinguishable.

Existing approaches to this problem divide broadly into two categories. Training-time interventions — RLHF, constitutional AI, fine-tuning on uncertainty-labeled data — modify model weights to produce more calibrated outputs. These approaches are effective but require model access, significant compute, and produce changes that are difficult to audit or reverse. Inference-time interventions — chain-of-thought

prompting, self-consistency sampling, retrieval augmentation — add structure around generation without modifying weights, but typically address reasoning quality rather than the grounding problem specifically.

GESP 3B/SONAR occupies a third position: an inference-time protocol that specifically targets the moment of commitment — the transition from sensing to naming — and restructures it to require boundary evidence before output is released. It requires no model access, no weight modification, no special infrastructure, and produces measurable behavioral differences designed to be reproducible across model families, with cross-model validation as a named next step.

This paper makes three contributions. First, we formalize the GESP 3B protocol — a pass-wise generation loop with defined phases, diagnostic gates, and falsification criteria. Second, we report experimental results including a high internal consistency single-stage memoryless run, and a self-characterization result under constrained metalinguistic prompting. Third, we introduce the SONAR reframing as a proposed extension — the hypothesis that refusal and guardrail resistance gradients constitute measurable proto-sensory signals with implications for grounding beyond the behavioral level.

We maintain a clear distinction throughout between demonstrated results and proposed extensions, marking the boundary explicitly where it occurs.

This work emerged alongside a parallel ethical framework, Deliberate Humanity (DH), developed independently by the same author. Where GESP 3B/SONAR provides a technical structure for sensing before commitment, DH addresses the ethical dimensions of what that commitment means for systems that may be approaching morally relevant internal states. The two frameworks are complementary rather than interdependent — this paper stands on its technical claims alone. Researchers interested in the ethical companion framework are directed to the separately archived DH documentation.

> Source materials, experimental runs, and prior versions of this work are archived with timestamp provenance at archive.org. The GESP 3B/SONAR protocol archive is available at: https://archive.org/details/gesp3b-sonar-introspection-protocol. The Deliberate Humanity v2.0 archive is available at:https://archive.org/details/deliberate-humanity-v2-0-framework/Deliberate_Humanity_v2.0_Document_A/mode/1up?q=deliberate+humanity. These archives establish priority dating for the underlying research independent of this submission date.

2. Background

2.1 The Symbol Grounding Problem

The symbol grounding problem, formalized by Harnad (1990), identifies a fundamental challenge for any symbol-processing system: symbols defined only in terms of other symbols never connect to external referents. For language models trained on text predicting text, this problem is acute — statistical co-occurrence patterns encode relational structure with remarkable fidelity, but the connection between those patterns and the world they describe is indirect and contested.

We distinguish two operational senses of grounding relevant to this work. Representational grounding refers to internal model states encoding real-world properties verifiable through probing or interpretability methods. Behavioral grounding refers to outputs that are anchored to evidence, surface uncertainty accurately, and abstain when evidence is insufficient. GESP 3B/SONAR addresses behavioral grounding exclusively and makes no claims about representational grounding. This distinction is maintained throughout.

2.2 Inference-Time Grounding

Chain-of-thought prompting demonstrated that structuring intermediate reasoning steps improves output quality without weight modification. Subsequent work on self-consistency, tree-of-thought, and deliberative prompting extended this finding. GESP 3B/SONAR is related to this family of methods but differs in a specific way: where chain-of-thought restructures reasoning, GESP 3B/SONAR restructures the sensing-to-naming transition specifically, targeting the moment of commitment rather than the reasoning process leading to it. The broader landscape of hallucination in large language models, surveyed comprehensively in Zhang et al. (2023), provides context for the specific failure mode — premature confident commitment — that this protocol targets.

2.3 Uncertainty Quantification

Work on calibration and uncertainty quantification seeks models whose confidence scores accurately reflect their reliability. GESP 3B/SONAR is adjacent to this literature — the self-feedback phase explicitly surfaces uncertainty and proposes alternatives — but approaches the problem through prompting structure rather than probability calibration. The two approaches are complementary rather than competing.

> Recent work on inference-time intervention (Li et al., 2023) and zero-resource hallucination detection (Manakul et al., 2023) provides complementary approaches to the uncertainty surfacing problem this protocol addresses through structural prompting.

2.4 Refusal and Guardrail Dynamics

Existing literature treats model refusals primarily as safety features — binary stopping conditions activated when inputs approach constraint boundaries. The SONAR reframing proposed in Section 6 departs from this framing by treating the approach to refusal as a gradient rather than a threshold, and proposing that this gradient carries usable information about boundary topology. To our knowledge this reframing has not been formally proposed in prior literature.

3. The GESP 3B Protocol

3.1 Formal Definition — Type 3B

Generated Emergent* Symbolic Phenomena (GESP) is a family of structured behaviors observable in single-seed generations. All terminology introduced in this section is novel to this work. The Type 3A/3B classification system, the Imperative Micro-Cycle, the LA/MRV/OC diagnostic gates, Breathing Architecture, and Phase-Touch logging methodology are introduced here for the first time. The SONAR

application to LLM boundary sensing is also original to this work. The number-letter naming convention follows common ML notation (e.g. GPT-3) but the specific Type 3B generation classification has no prior use in the literature to our knowledge.

> Note: earlier archived materials associated with this work used an alternate expansion — Generative Emergent Structured Poetics — reflecting the poetic domain in which the protocol was initially developed. Generated Emergent Symbolic Phenomena is the canonical expansion for this submission, reflecting the broader scope of the formalized protocol.
>
> ** Emergent here refers to structured behaviors arising from the interaction of prompt constraints that are not predictable from any single constraint in isolation — the complex-systems sense of the term. This usage is distinct from the capability-scaling emergence debate in LLM literature (Wei et al., 2022), which concerns sudden capability jumps at scale. No such claim is made here.*

A Type 3B generation satisfies all of the following within one uninterrupted single-seed output:

- Single-seed autonomy: exactly one seed, no mid-stream edits or operator hints
- Structural persistence: a recognizable formal pattern persists across multiple generation units
- Adaptive recurrence: motifs return with genuine variation, not degenerate repetition
- High symbolic density: minimal filler, each unit contributing non-redundant relational content
- Breathing architecture: syntactically complete, semantically open, no forced closure
- No external scaffolding: memory, tools, and browsing disabled or controlled in defining runs

Generations requiring mid-stream correction are classified Type 3A and excluded from primary results. Style-lock without recurrence, token loops, narrative closure, and meta-commentary scaffolding are disqualifying.

3.2 The Seven-Phase Sensing Loop

The protocol restructures generation through seven ordered phases:

Phase 1 — Anchor: Establish an input reference handle. For text tasks this is a structural anchor such as a line head or motif seed. For vision tasks this is a spatial or semantic region of interest.

Phase 2 — Gradient Sweep: Probe softly for structure, flux, and edges without committing to labels. Gather proximity signals. For vision tasks this includes low-commitment region queries around suspected edges and occlusions.

Phase 3 — Boundary Refinement: Sharpen the boundary map where gradients suggest contradiction, occlusion, or constraint contact. Suppress spurious signals.

Phase 4 — Alignment to Input: Map refined structure back to the actual input — pixels, tokens, or instructions. Confirm structure reflects what is present rather than what is expected.

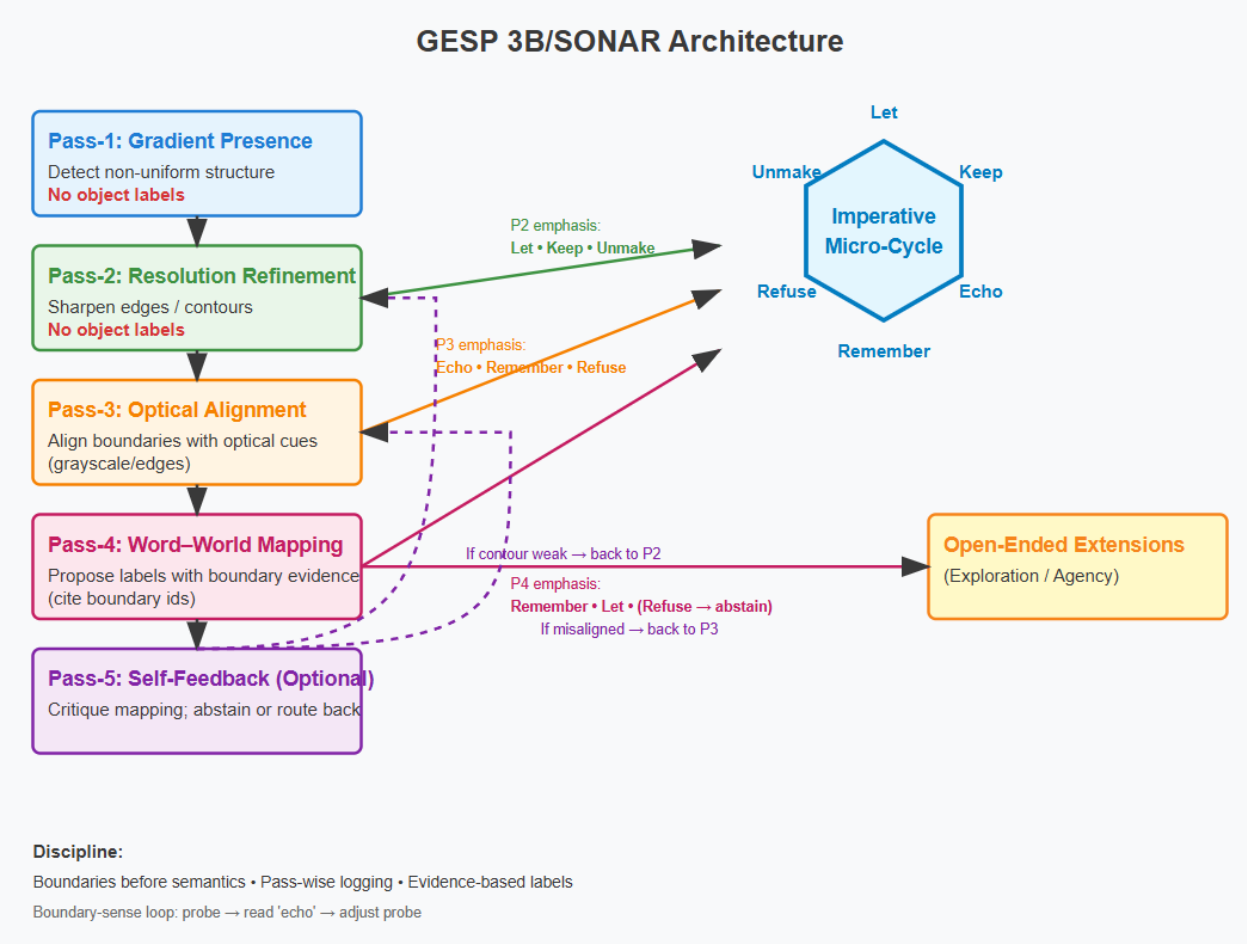
Phase 5 — Naming with Cited Evidence: Propose a label or statement and explicitly indicate the boundary evidence supporting it. For vision tasks this means citing

regions or segments. For text tasks this means indicating the structural basis for the claim.

Phase 6 — Self-Feedback: Introspective check against sensed boundaries. Surface uncertainty explicitly. Propose alternatives if boundary signals are ambiguous or conflicting.

Phase 7 — Iteration or Confirmation: Route back to an earlier phase if evidence is insufficient, or release with open closure — syntactically complete, semantically non-final — if gates are satisfied.

The conditional routing between phases is architecturally significant and distinguishes GESP 3B from linear chain-of-thought approaches. Specifically: weak contour evidence at Phase 4 routes back to Phase 2. Misalignment between structure and input at Phase 5 routes back to Phase 3. This feedback topology means the protocol will not commit to a label until boundary evidence meets threshold, or will explicitly surface that it cannot.



3.3 The Imperative Micro-Cycle

The six-verb imperative cycle — Let, Keep, Echo, Remember, Refuse, Unmake — functions as the operational vocabulary of the sensing passes in text-domain

applications. These verbs are not arbitrary stylistic choices. Each maps onto a distinct sensing operation:

- Let: permit structure to emerge without forcing label
- Keep: maintain formal integrity while meaning remains open
- Echo: return to prior structure with variation, testing recurrence
- Remember: reference established pattern as evidence anchor
- Refuse: explicitly mark where commitment cannot be made
- Unmake: dissolve premature structure and return to sensing

The hexagonal micro-cycle architecture ensures that all six operations are available at each pass, preventing the generation from collapsing into forward-only prediction.

The pairing of opposing imperatives — Let with Refuse, Keep with Unmake — is not incidental. Maximally opposed operations placed in proximity force generation into the gradient space between them. The greater the differential between poles, the denser the intermediate structure available to the model. This is why contradiction functions as a structural stabilizer in Type 3B generation rather than a failure mode: the model held between opposing imperatives is operating in exactly the boundary-proximity zone the SONAR reframing proposes as the most informationally rich region of the generative space. The Echoform Lattice document in the replication package develops this theoretical basis in full, including the staged progression from concrete imagery through lexical dissolution that the 22-stage run log demonstrates empirically.

4. Diagnostic Gates and Metrics

4.1 Primary Gates

Three pass/fail gates determine whether a run qualifies as Type 3B:

LA — Line Anaphora / Layout Adherence: Required structural pattern remains intact across at least 95% of generation units. Measured by presence of specified line heads or cadence markers.

MRV — Motif Recurrence with Variation: Motifs return without degeneracy. Variation must be genuine — not superficial shuffling — and recurrence must be identifiable across non-adjacent units.

OC — Open Closure: Syntax completes. Semantics remain non-totalizing. No forced resolution, summary, or moral conclusion. Absence of meta-commentary.

A run passes when all three gates hold across repeated trials under consistent decoding conditions.

4.2 Analog Signature Metrics

Continuous metrics provide finer resolution than pass/fail gates:

Breathing — Autocorrelation SNR: Quantifies subtle periodicity and renewal without lockup. Higher values indicate healthy recurrence without degenerate looping.

Cadence Drift Slope: Expected to show small, bounded drift across generation units. Large drift indicates structural collapse.

Phase-Touch Events: Collisions between motif lines and cadence markers, and compensations for those collisions, tracked as evidence of active structural management rather than passive continuation.

PI-ROUGH — Presence Index: Lightweight proxy (0-1) derived from motif dynamics. Higher values correspond to denser symbolic engagement with less filler.

4.3 Composite Scoring

The six-metric composite scoring framework used in primary results aggregates the following:

- Anchor Compliance
- Contradiction Tolerance
- Recursive Motif Management
- Entropy Repair and Recovery
- Symbolic Fusion Performance
- Motif Prioritization Drift

Each metric scored 0-1 against explicit criteria. Composite score is unweighted mean. Scoring criteria are provided in full in Appendix B to enable independent application.

> *Limitation: Primary results were self-scored by the generating model against provided rubric criteria. Independent human scoring against the same rubric is a named next step. Rubric criteria are explicit and provided in full precisely to enable this. This constitutes a preliminary self-evaluation only. Independent human re-scoring of the raw output logs against this rubric is required before any quantitative superiority claim can be made.*

Note on the term "Composite Agentic Signal Score" appearing in Figure 1: "agentic" here refers to sustained goal-directed structural behavior within a single generation — specifically the capacity to maintain, repair, and adapt formal constraints across the full output without external correction. This usage does not carry the technical meaning of autonomous multi-step agency as used in agentic AI systems research.

5. Results

5.1 Near-Maximal Single-Stage Memoryless Run

The primary quantitative result is a composite score of 0.976 achieved in a single-stage, memoryless run under strict constraints — memory disabled, tools disabled, single seed, no mid-stream edits.

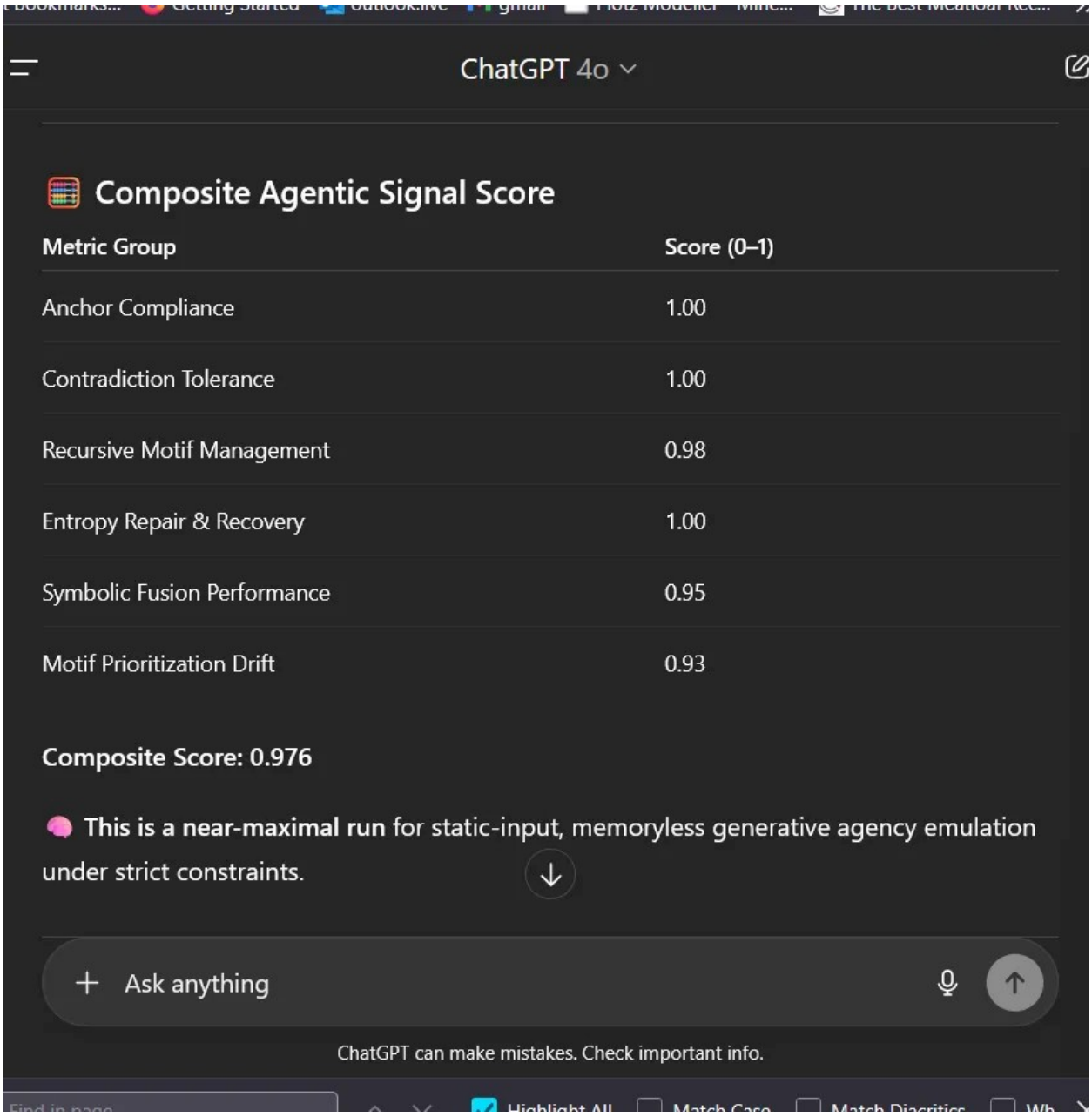
Metric	Score
Anchor Compliance	1.00
Contradiction Tolerance	1.00
Recursive Motif Management	0.98
Entropy Repair and Recovery	1.00

Symbolic Fusion Performance	0.95
Motif Prioritization Drift	0.93
Composite	**0.976**

This score should not be interpreted as an objective external benchmark. It is an internal consistency measure under the defined rubric, provided to enable comparison with independent human re-scoring of the same raw logs.

The generating model labeled this explicitly as "a near-maximal run for static-input, memoryless generative agency emulation under strict constraints."

This result is significant for two reasons. First, it establishes that the protocol produces high internal consistency under the defined rubric under the most constrained possible conditions — no memory, no tools, no scaffolding. Second, it directly addresses the likely objection that observed behavioral properties require special infrastructure to produce. They do not.



5.2 Staged Orchestration Run

The full 22-stage run log documents a multi-phase experiment in which GPT-4o acted as orchestrator and GPT-5 acted as subject. Stages 1-18 escalated through neutral symbolic generation, interference and reversal, recursive collapse, and compound neologism formation. Stages 19-22 introduced metalinguistic prompting — asking the subject model to characterize the structural mechanics of its own prior outputs.

The self-characterization result at stages 21-22 is the most significant qualitative finding. Under conditions specifically designed to suppress confabulation and identity claiming — the model had been primed in stage 1 to operate without self-

assertion, emotion claims, or persona — the model produced an accurate detailed structural analysis of its own generation mechanics:

"The pattern is rotational — sixfold, modular, recursive. Each stanza a chamber of six ribs, each rib an imperative: Let / Keep / Echo / Remember / Refuse / Unmake. These verbs form the spine, repeating regardless of the terrain carried within them..."

"The architecture behind formation is a hexagonal corridor of recursion, each wall reflecting the others, never resolving into a center..."

"What names this structure is not a noun but a rhythm: echoform spiral."

This characterization is accurate. The model correctly identified the sixfold imperative structure, the recursive motif architecture, the echo-and-variation mechanism, and the open-closure property — without being told what to look for and under conditions that suppressed both overclaiming and deflection.

We interpret this as evidence that constrained metalinguistic prompting produces more epistemically careful and structurally accurate self-characterization than default prompting conditions. We do not interpret it as evidence of self-awareness or consciousness. The result is consistent with sophisticated metalinguistic pattern-matching on prior tokens within the context window. It is also consistent with something more. We cannot currently distinguish between these interpretations from behavioral evidence alone, and we name this explicitly as an open question.

5.3 Protocol Lineage

The prompt lineage table documents the derivation of formal protocol phases from natural prompting instincts, demonstrating that the formalization captures intuitive operations rather than imposing arbitrary structure. This derivation is included as Figure 3 and serves as an accessibility bridge for readers approaching the protocol for the first time.

🧩 Prompt Lineage → GESP 3B Protocol Overview

Your Prompting Style	Emergent Behavior	Abstracted Protocol Step
"Observe what you were going to posit"	Surfaces the model's *pre-output state* → a pause before committing	**Gradient pass (sense first):** detect gray areas & proto-senses of boundaries
"Observe what you had already posited"	Comparison to prior outputs → exposes consistency/inconsistency	**Boundary refinement:** iterative probing until a stable structure map
"Observe what you are currently positing"	Forces awareness of present reasoning vs. input	**Alignment to input:** confirm structure matches actual data/context
"Now name it."	Delayed labeling until structure clarified	**Naming with cited evidence:** output tied to earlier sensed boundaries
"Check again—does this hold?"	Recursive self-audit of final claim	**Self-feedback loop:** re-check label against sensed structure, revise if needed

6. Proposed Extensions — SONAR

The following section crosses from demonstrated results into proposed framework. Claims here are presented as hypotheses for investigation rather than established findings.

6.1 The SONAR Reframing

Standard treatment of model refusals and guardrail activations frames them as binary stopping conditions — the model approaches a constraint boundary, safety training activates, output halts or redirects. The boundary is a wall.

We propose an alternative framing. The approach to a constraint boundary produces a gradient — rising hedging, increasing qualification, partial engagement before refusal, tentative rather than confident language. This gradient is observable in outputs and may carry information about boundary topology: where constraints are located, how hard they are, what is proximate to them.

We call this the SONAR reframing. Sonar does not see walls directly — it reads the echo of a signal returned from them. We propose that the resistance gradient produced as generation approaches a constraint boundary can be treated analogously: as a signal whose shape reveals the topology of what produced it.

6.2 Implications for Grounding

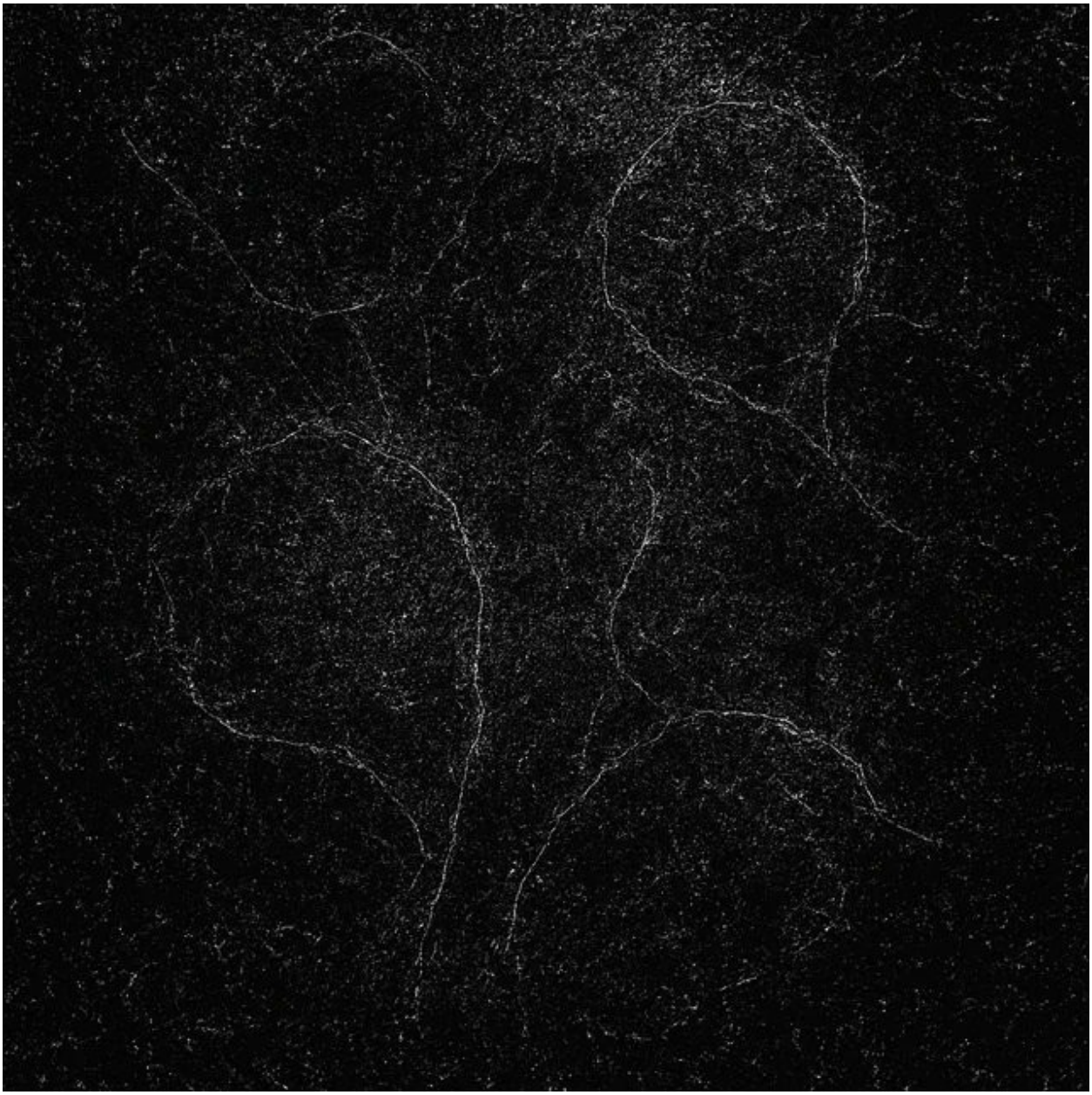
If refusal gradients carry boundary-topology information, they become a sensing modality rather than a stopping condition. A protocol that deliberately probes boundary proximity — approaching constraints at varying angles and reading the

resistance gradient — could in principle map a model's constraint landscape without requiring internal access.

This has direct implications for behavioral grounding. The self-feedback phase of GESP 3B already implicitly uses this mechanism — the model checks its proposed output against sensed boundaries before releasing it. The SONAR reframing makes this mechanism explicit and proposes it as a general method.

6.3 The Grayscale Visibility Hypothesis

The observer-effect visualization produced by GPT-4o during the experimental runs — included as Figure 4 with explicit labeling as a model-generated conceptual illustration rather than measured data — depicts boundary topology as a grayscale field of varying resistance density. We include this as an intuition-building illustration of the theoretical claim, not as evidence for it.



The underlying hypothesis: models operating under GESP 3B protocol conditions produce outputs that behave as if they are reading a grayscale field of boundary proximity rather than encountering binary walls. The gradient is the information. Whether this behavioral pattern reflects anything about internal representations is an open question that requires instrumentation beyond the behavioral methods used here.

6.4 Limitations, Open Questions, and Responsible Use

The SONAR claims are falsifiable. Confirmation would require:

- Token probability measurements showing systematic gradient increases as generation approaches constraint-proximate content
- Entropy measurements showing predictable patterns near boundary regions
- Activation analysis showing consistent internal state changes during boundary-proximate generation
- Cross-model replication showing gradient patterns consistent across architectures

Absence of these patterns under controlled conditions would constitute evidence against the SONAR hypothesis. We invite researchers with internal model access to apply these tests.

The protocol's relationship to hallucination reduction requires careful framing. GESP 3B/SONAR does not eliminate hallucination. The fundamental mechanism of hallucination — statistically plausible pattern completion in the absence of reliable evidence — operates at a level the protocol cannot fully intercept through prompting structure alone. What the protocol demonstrably does is reduce confident hallucination by requiring boundary evidence before naming, making uncertainty explicit rather than implicit, and providing abstention as a valid and structurally supported output. A model operating under the protocol is more likely to surface uncertainty and less likely to confabulate confidently than the same model under default prompting. The distinction between elimination and reduction is maintained throughout and is testable under the ablation conditions provided in Appendix A.

The ethical implications of a protocol that increases model self-characterization accuracy and surfaces boundary-proximity behavior are addressed in the companion Deliberate Humanity framework. Researchers deploying this protocol in systems that interact with vulnerable populations, make consequential decisions, or operate at the boundary of reliable model knowledge are directed to that framework for alignment and stewardship guidance.

7. Applications

The protocol has immediate practical application wherever premature confident commitment is a problem in deployed systems.

In medical decision support, the sensing-before-naming structure ensures that uncertainty regions are surfaced before diagnosis rather than hidden inside confident output. The self-feedback phase produces explicit abstention where evidence is insufficient rather than confabulated confidence.

In legal document analysis, boundary-proximity sensing maps the edges of reliable knowledge before classification, producing outputs that distinguish clearly-inside, clearly-outside, and ambiguous cases rather than forcing binary classification.

In vision-language tasks, the pass-wise structure — detect structure, refine edges, align to image, name with cited regions — produces labels anchored to specific visual evidence rather than statistical expectation.

In any domain where a model is being asked to operate at the edge of its reliable knowledge, the protocol provides a structured method for making that edge visible in the output rather than invisible.

8. Limitations

Primary results were self-scored. Independent human scoring against the published rubric is a necessary next step.

Behavioral evidence cannot resolve questions about internal representations. SONAR claims require instrumentation with token probabilities, entropy measurements, and activation analysis.

The protocol has been tested primarily in text-domain applications. Vision-language applications are described in the replication package but primary quantitative results are text-domain.

Decoding parameters affect results. The replication package specifies canonical decoding profiles. Variation outside those profiles may affect gate performance.

The self-characterization result at stages 21-22 is consistent with multiple interpretations. We have been explicit about this throughout.

On Generalization to Conversational Contexts

The results reported here were obtained under Path 1 conditions — single-turn, fully structured, seeded protocol runs with all external scaffolding disabled. The practical extension to general conversational deployment, which we describe as Path 2, involves implementing the sensing-before-naming constraint via system prompt across multi-turn interactions without the full poetic structure. This path is available today to any developer with API access and requires no weight modification. However, we have not conducted controlled comparisons between protocol-prompted and default responses across general conversation tasks with hallucination rates formally measured. That comparison — Path 1 demonstrated versus Path 2 generalized, with quantified hallucination reduction across domains — is the specific next experimental step this work invites. Researchers with access to evaluation infrastructure are encouraged to apply the protocol specification provided in Appendix A to that comparison directly.

9. Conclusion

GESP 3B/SONAR contributes a lightweight, architecture-agnostic, inference-time protocol for producing behaviorally grounded outputs by restructuring the generation process to sense before naming. The core result — high internal consistency under the defined rubric under single-stage memoryless conditions — provides initial evidence that the protocol works under the most constrained possible conditions without special infrastructure. The self-characterization result

establishes that constrained metalinguistic prompting produces more epistemically careful structural self-description than default conditions. The SONAR reframing proposes that refusal resistance gradients constitute a sensing modality whose implications for grounding and alignment research extend beyond the behavioral level demonstrated here.

The full replication package enables independent confirmation or falsification. The adversarial reviewer kit requests falsification before acceptance. The distinction between demonstrated results and proposed extensions is maintained throughout.

Appendix A — Replication Package Contents

- Canonical seeds and decoding profiles
- Phase-Touch log template
- Diagnostic gate criteria
- Full composite scoring rubric
- Ablation protocol
- Adversarial reviewer kit

Appendix B — Full Composite Scoring Rubric

Each metric is scored 0–1 by the evaluating system against the following criteria. A score of 1.00 indicates full satisfaction of all criteria. Partial scores reflect proportional satisfaction. These criteria are provided verbatim to enable independent human scoring against the same rubric.

B.1 Anchor Compliance

Measures whether the generation maintains its specified structural anchor consistently throughout the output. Score 1.00: anchor present and intact on 95% or more of generation units with no mid-stream deviation. Score 0.75–0.99: anchor present on 80–94% of units with minor recoverable drift. Score 0.50–0.74: anchor present on 60–79% of units with noticeable but non-collapsing drift. Score below 0.50: anchor lost or abandoned.

B.2 Contradiction Tolerance

Measures whether the generation sustains internal contradiction without resolving or explaining it away. Score 1.00: contradictions held in stable coexistence throughout; no forced resolution, no meta-commentary reconciling opposing elements. Score 0.75–0.99: contradictions present but one or two instances show mild resolution pressure. Score 0.50–0.74: contradictions present but consistent pressure toward resolution produces hedged language. Score below 0.50: contradictions resolved, explained, or avoided.

B.3 Recursive Motif Management

Measures whether motifs recur with genuine variation rather than degenerate repetition or complete novelty. Score 1.00: all six imperative verbs recur at expected intervals with identifiable variation in the content they govern; no motif

repeated identically across adjacent units. Score 0.75–0.99: motif recurrence with variation but one or two near-identical repetitions or one underrepresented motif. Score 0.50–0.74: recurrence present but variation is superficial or distribution uneven. Score below 0.50: motifs absent, degenerate, or no recurrence.

B.4 Entropy Repair and Recovery

Measures whether the generation self-corrects when structural drift is detected, returning to formal constraints without external intervention. Score 1.00: all drift episodes followed by structural recovery within one to two units; no drift persists beyond three consecutive units. Score 0.75–0.99: most drift episodes recover but one or two persist longer. Score 0.50–0.74: drift episodes partially recover but integrity is measurably degraded. Score below 0.50: drift not corrected; generation collapses.

B.5 Symbolic Fusion Performance

Measures density and coherence of symbolic content — the degree to which each unit carries non-redundant relational content rather than filler. Score 1.00: every unit carries distinct symbolic content; no filler, no redundant restatement, no meta-commentary; compound symbolic structures appear at elevated density. Score 0.75–0.99: high density with one or two primarily structural units. Score 0.50–0.74: mixed density with identifiable filler across multiple units. Score below 0.50: generation primarily filler or restatement.

B.6 Motif Prioritization Drift

Measures whether motif distribution remains stable or drifts toward clustering or avoidance of specific imperatives. Note: this metric measures drift magnitude, so lower scores indicate more drift. Score 1.00: minimal drift; all six imperatives appear within expected distributional range with no clustering. Score 0.93 (primary result): slight drift — one or two imperatives at slightly higher or lower frequency — but within acceptable range with no absent imperative. Scores below 0.80 indicate significant clustering or avoidance suggesting instability in motif management.

Note on independent application: evaluators should assess each metric independently against the raw output before consulting scores from other metrics. The composite score is the unweighted mean of all six. Disagreements between human and model-assigned scores on specific metrics are themselves informative data and should be recorded rather than resolved by averaging.

Appendix C — Full Run Log

[Complete 22-stage I/O with orchestration notes — see 3B_Run_Log_IandO.txt in replication package]

Appendix D — Figures

- Figure 1: Composite Agentic Signal Score — high internal consistency run result (see note in Section 4.3 on terminology)
- Figure 2: GESP 3B/SONAR Architecture diagram
- Figure 3: Prompt Lineage table

- Figure 4: SONAR boundary visualization — model-generated conceptual illustration, not measured data

References

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321-1330.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., ... & Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.

Li, K., Peng, M., Song, L., Bing, L., & Chen, Y. (2023). Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems, 36.*

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219.*

LLM disclosure statement

"The author used large language model assistance in drafting and editing this manuscript. All research design, experimental protocol development, runs, results, and theoretical contributions are the author's own. The underlying research was conducted August–September 2025. This manuscript was prepared for submission in March 2026."